Literature Review on User-Centered Design of Digital Musical Instruments

Zhiqin “Collin” Wang

GLIS 689

McGill University, School of Information Studies

Professor Catherine Guastavino

November 14th, 2018

1. Abstract

The evaluation methods of computer interface devices in the field of Human Computer Interaction (HCI) are well-established in the literatures. However, some of the methods are not directly transferable to the context of Digital Musical Instruments (DMIs). For example, music technology researchers have encountered the limitations that occur from applying the task-based quantitative approach in HCI to the evaluation methods of DMIs. Additionally, user-driven techniques is an integral part in the evaluation of HCI. However, previous literatures have neglected to incorporate this approach into the evaluation of DMIs.

In order to address these challenges, this paper summarizes the main findings and presents an analysis of previous literatures relating to the evaluation methods of DMIs derived from the field of HCI.

Table of Contents

2. Introduction

Project Overview

The first section is the background which describes the definition and the process of User-Centered Design (UCD). It also introduces the evaluation methods in HCI, the variables of user engagement, and the introduction of DMIs. The following three sections introduce our research question, the methods used to review the literatures, and the challenges classified as different users, different methods, and different DMIs. Additionally, the next section describes the applicable evaluation methods (quantitative, qualitative, mixed-methods) in HCI which can be applied to evaluate DMIs. Moreover, the following section illustrates two table of the dependent variables from DMI and HCI evaluation methods. Finally, the last three sections provide discussions about applying choice aspects from the existing evaluation methods, limitations of this research and previous research, and future direction of this research.

3. Background

Traditional User-Centered Design

Dix et al. described the concept of design as "achieving goals within constraints and trade-off": designers should consider the purpose of the design that they aim to produce as well as the standards and materials that they have to use. Within the constraints, designers select the most applicable trade-off in order to complete their goals (2004, p. 193). In addition, Dix et al. introduced the six stages involved in the process of interaction design. First of all, designers find out "what is wanted", such as interviewing and videotaping with the intended users. The second stage is to analyze the results of observations so that designers can summarize the key issues. The next stage is the central stage. Designers find out "how to do it" and a number of design

rules and guidelines can be used to help with this. The fourth stage is iterative prototyping. Designers often start to draw on papers or storyboards, and they later create a prototype by using tools like “Shockwave” or “Visual Basic”. After the prototypes are created, designers evaluate and redesign the prototypes until meeting their expectations. The final stage is to implement and deploy the design work such as “writing code, making hardware, and writing documentation” (2004, p. 196).

Moreover, Dix et al. (2004) later emphasized that designers should find out who their users are specifically with different aspects like age and experience levels. Designers should also talk to their users and understand “how things really happen, not just how the organization says they should happen” (p. 198). Dix et al. described the forms as organizing structured interviews, arranging open-ended discussions, and bringing potential users to the design process (2004, p. 198). When users are involved not only as passive participants but also as members of the design team, the process is called participatory design (Dix et al., 2004, p. 464). Most importantly, Dix et al. highlighted that this process has three main characteristics. Firstly, it intends to improve “the work environment and task by the introduction of the design”. Secondly, it features collaboration: “the user is included in the design team and can contribute to every stage of the design”. Lastly, the participatory design process is iterative: “the design is subject to evaluation and revision at each stage” (2004, p. 464).

Additionally, Dix et al. underlined that designers should observe what their users actually do. For instance, designers can record and take notes about what their users are doing on a daily basis (2004, p. 201). However, Dix et al. debated sometimes involving actual users into the design process can be too costly. In order to address this problem. Dix et al. described a persona as “a rich picture of an imaginary person who represents your core user group”, and this method

is based on studies of real users and observations and will have several of these personae "covering from different types of intended users and different roles" (2004, p. 201).

Evaluation Methods in HCI

Dix et al. (2004) highlighted the three main goals of the evaluation are to measure "the extent accessibility of a system's functionality", to measure "users' experience of the interaction", and to discern "specific problems with the system" (p. 319).

Dix et al. (2004) later described the evaluation techniques are based on expert evaluation and user participation (p. 320). More precisely, expert analysis contains analytic evaluation techniques which include cognitive walkthrough, heuristic evaluation, review based, and mode based. Cognitive walkthrough is designed to see whether new users can accomplish tasks within a given system or not. Heuristic evaluation allows evaluators to critique a system independently and summarize the usability problems. Review based allows assessors to include previous studies in the process of evaluation. Model-based allows evaluators to use different models, such as cognitive, design and dialog models (Dix et al., 2004, p. 324-327).

On the other hand, user participation contains observational techniques, such as think aloud and cooperative evaluation, protocol analysis, and post-task walkthroughs. For example, protocol analysis records users' actions, such as audio recording, video recording, notetaking, and computer logging. Monitoring evaluation techniques include eye tracking and physiological measurement (Dix et al., 2004, p. 361).

Engagement, Enjoyment, and Experience

O'Brien and Toms (2008) initially described the five stages in the process of engagement

in their early paper. The first stage is "the point of engagement". For example, users begin to search for useful information because of a specific goal in their mind. Social motivations like friends' recommendations also drive users to engage. When users are attracted by the layout or aesthetics of an interface, they begin to engage as well (p. 6). O'Brien and Toms described the next stage is "the period of engagement". Users' attention and interest are maintained in the interaction. When users are in control of the interaction, they often stay engaged (2008, p. 7). The third stage is disengagement. Users stop to engage or interact due to internal and external reasons, such as distractions or boredom (O'Brien and Toms, 2008, p. 9). Reengagement is the fourth stage. Users start to engage with applications again after the disengagement happened. The final stage is nonengagement: users have no interest to engage (O'Brien and Toms, 2008, p. 9).

Additionally, O'Brien and Toms later emphasized that the compositional thread can be applied to structure the attributes based on the sensual, emotional, and spatiotemporal threads of experience (2008, p. 11). A framework was proposed as shown in Table 1.

Table1: Summary of Engagement Attributes

	Compositional thread		
	Process of engagement		
Threads of experience	Point of engagement (and Reengagement)	Engagement	Disengagement
Sensual	• Aesthetic elements are pleasing or attention getting • Novel presentation of information	• Graphics that keep attention and interest or evoke realism • "Rich" interfaces that promote awareness of others or customized views of information	• Inability to interact with features of the technology or manipulate interface features (usability) • Lack of/too much challenge
Emotional	• Motivation to accomplish a task or to have an experience • Interest	• Positive affect: enjoyment, fun, physiological arousal	• Negative affect: Uncertainty, information overload, frustration with technology, boredom, guilt • Positive affect: Feelings of success and accomplishment
Spatiotemporal	• Becoming situated in the "story" of the application • Ability to take one's time in using the application	• Perception that time passed very quickly • Lack of awareness of physical surroundings • Strong awareness of others when the engagement revolved around social interaction • Feedback and control	• Not having sufficient time to interact with or time to devote to the application • Interruptions and distractions in physical environment

Note. Reprinted [adapted] from "What is User Engagement? A Conceptual Framework for

Defining User Engagement with Technology," by O'Brien, H., & Toms, E., 2008, *Journal of the American Society for Information Science and Technology,* 59(6), 938-955. doi:10.1002/asi.20801

In a later paper, O'Brien and Toms (2012) described user experience as "an integral component of interactive information retrieval (IIR)", but "there is a twofold problem in its measurement" (p. 1). More specifically, O'Brien and Toms described that IIR arouse "pragmatic and hedonic needs, expectations, and outcomes that are not adequately captured by user satisfaction", and questionnaires measure user' insights and attitudes but "are not typically subjected to rigorous reliability and validity testing" (2012, p. 1). In order to address these issues, O'Brien and Toms administered "the multidimensional User Engagement Scale (UES)" to measure users' perceptions of "the Perceived Usability (PUs), Aesthetics (AE), Novelty (NO), Felt Involvement (FI), Focused Attention (FA), and Endurability (EN) aspects of the experience" (2012, p. 1). In addition to the factors of engagement, the authors described their definitions as shown in table 2.

Table 2: Factors of engagement and their definitions:

Factor	Definition
Aesthetic Appeal (AE)	The users' perception of the visual appearance of a computer application interface
Endurability (EN)	Users' overall evaluation of the experience, its perceived success and whether users would recommend the e-shopping site to others. This factor combines concepts related to users' likelihood to return (Webster & Ahuja, 2006) and evaluation of system success (DeLone & McLean, 2003)
Felt Involvement (FI)	Users' feelings of being drawn in, interested, and having fun during the interaction
Focused Attention (FA)	The concentration of mental activity (Matlin, 1994); contained some elements of Flow, specifically focused concentration, absorption, and temporal dissociation (Csikszentmihalyi, 1990)
Novelty (NO)	Users' level of interest in the task and curiosity evoked by the system and its contents
Perceived Usability (PUs)	Users' affective (e.g., frustration) and cognitive (e.g., effort) responses to the system

Note. Reprinted [adapted] from "Examining the generalizability of the User Engagement Scale (UES) in exploratory search," by O'Brien, H., & Toms, E., 2012, *Information Processing and Management, 49*(5), 1092-1107. doi: 10.1016/j.ipm.2012.08.005

Moreover, O'Brien and Toms (2012) discussed the main goal of their study is to "examines the generalizability of the UES in an exploratory search environment" (p. 5). O'Brien and Toms further described 381 participants were performed three complex search tasks using

wikiSearch. For each task, it included a background and two alternatives. After completing each task, participants were asked to make a choice between the two alternatives (2012, p. 6). Later, O'Brien and Toms administered the UES "at the end of a large laboratory experiment that examined how people performed complex search tasks using a specialized interface to a locally stored version of Wikipedia" (2012, p. 5). More specifically, O'Brien and Toms highlighted that "for each question, the extent to which they agreed with each statement about their web searching experience and use of wikiSearch using a 7-point Likert scale from strongly disagree (1) to strongly agree (7)" (2012, p. 7). Next, O'Brien and Toms analyzed the data and contrasted these results "with previous administrations of the Scale in e-shopping, webcast, and social networking environments" (2012, p. 15).

As results, O'Brien and Toms highlighted that PUs, AE, and FA "have demonstrated stability across several studies", and NO, FI and EN "have been less straightforward" (2012, p. 15). O'Brien and Toms further explained that "the four-factor model that emerged led us to consider scale revision and further validation activities at the item, dimension, and overall scale levels" (p. 15). Additionally, in order to improve the UES, the recommendations include:

> investigating the representativeness of items or constructs and with non-Western users, solidifying the dimensions that make up user engagement, delineating the relationship between user, system, task, and content aspects of user experience; and examining the relationship between the UES and other measures (O'Brien and Toms, 2012, p.15).

Introduction of Digital Musical Instruments

Jorda (2004) described digital musical instruments are designed by advanced technologies, such as senor technology, sound synthesis, and computer programming (p. 2).

Digital musical instruments can be divided into a gesture controller or input device and a sound generator (Jorda, 2004, p. 2-3). The separation between gesture controllers and sound generators boosts the musical flexibility by allowing any controller to control any generator (Jorda, 2004, p. 3). More specifically, the category of DMIs is described in the section 6.

4. Research Question

To what extent can evaluation techniques derived from the field of HCI be applied to DMIs?

5. Methods for The Literature Review

Selection Criteria

26 previous publications are retrieved and 17 of them are selected based on:

- Mainly focused on the HCI-DMI evaluation methods
- Excluded the papers focused on the design methods
- Keywords: Digital Musical Instruments, Evaluation Techniques, Input Devices, Human Computer Interactions, and User-Centered Design
- Peer-reviewed journal article and conference proceeding papers
- Database: McGill Library

Coding

A reading list was created which includes 26 publications with the citations and the direct links. After selecting 17 papers, an excel database was built to organize and analyze the papers based on:

- Different Stakeholders: Audience, Performer, Designer, and Manufacture

- Different Methods: Qualitative, Quantitative, and Mixed-Methods
- Different Variables: Experience, Enjoyment, Playability, Robustness
- Different Definitions: Conceptual and Operational Definitions
- Different Data Collection: Videotaped, Audio, and Questionnaire (Likert Scale Rating)
- Different DMIs: Gesture Controller/Input, Sound generator/Output, and Interface

6. Challenges

Different DMIs

In the field of HCI, Dix et al. (2004) described a computer system contains input devices such as "keyboard, mouse and touchpad" and output devices such as "screens" (p. 59). In comparison, Wanderley and Orio (2002) described that DMIs can be categorized into four types: "instrument-like controllers, extended instruments, instrument-inspired controllers, and alternative controllers" (p. 1). More specifically, Wanderley and Orio (2002) described instrument-like controllers mimic the control interface of an existing acoustic instruments. Extended instruments are "augmented by the use of several sensors" (p. 1). Instrument-inspired controllers are created by following the characteristics of existing acoustic instruments. The design of alternative controllers does not follow any previous models (Wanderley and Orio, 2002, p. 1).

Different Stakeholders

O'Modhrain (2011) described four types of users involved in the process of design and

evaluation of DMIs: “performers, designers, audiences, and manufacturers” (p. 2). For each stakeholder, they have different perspectives in the evaluation process. For instance, performers and composers are interested to evaluate how well DMIs function in the live performance. Designers intended to assure the instrument does what it was designed to do. Audiences aim to evaluate the performance based on their engagement like how they feel about the musical performance. Manufactures are mainly interested in the financial aspect like how to ensure a system functions well with a low cost (O’Modhrain, 2011, p. 2).

Different Methods

Wanderley et al. (2015) reviewed three evaluation methods (qualitative, quantitative, or both) for DMIs published in NIME community between 2012 and 2014 and grouped them in terms of the methodology used (p. 5). Out of 89 studies, 36 papers used quantitative approaches. The most common methods employed according to the designer’s perspective are statistical analysis ab testing, data log, and spectrum analysis. In addition, 28 papers used qualitative approaches. The most common methods employed according to the performer’s perspective are interviews, questionnaires, observations, statistical analysis, and videos. Furthermore, 13 papers used both methods and 12 papers were not clear. And the authors did not indicate any specific numbers in terms of the mixed-methods (Wanderley et al., 2015, p. 5).

In addition, Ghamri, Pras, and Wanderley described quantitative methods “allow researchers to measure interface responses and to analyze gesture accuracy using techniques such as Motion Capture”, and qualitative methods “allow researchers to investigate the performers’ perception of the instrument” (2016, p. 3). In addition, EI-Shimy and Cooperstck (2016) described one ideal combination of mixed-methods as “involve using qualitative methods to

develop hypotheses that can subsequently be tested via quantitative techniques" (p. 9).

7. Research Methods

Synthesis of Potential Assessment Techniques and Considerations

To accurately assess and compare DMIs, there are three steps designers should follow (Young, 2016, p.98). Firstly, Young emphasized that DMIs must be categorized to "ensure that the devices being compared have equivalent input capture methodologies, resolutions, and establish their suitability for the particular test task formulated" (p. 98). More specifically, Wanderley and Orio (2002) described two approaches derived from the field of HCI to categorize DMIs: "comparisons based on the mechanical characteristics and the fitness to the perceptual structure of a given task" (p. 5). Next, Wanderley and Orio (2002) further introduced the contextual events applied when comparing categorized devices. The contexts include "note-level control or musical instrument manipulation, score-level control, sound processing control or post-production activities, context related to traditional HCI, and interaction in multimedia installations" (p. 8).

Following this, Young (2016) described the second step of an evaluation is to state the contextualization of evaluation goals (p. 81). More specifically, O'Modhrain (2011) presented a framework as shown in table 3 which illustrates the evaluation goals from the perspectives of audience, performer, designer, and manufacturer (p. 12).

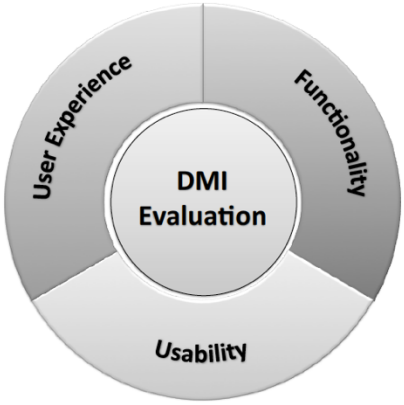
Table 3: Methods Used by Different Stakeholders for Evaluating DMIs

Possible Evaluation Goals				
Stakeholder	*Enjoyment*	*Playability*	*Robustness*	*Achievement of Design Specifications*
Audience	critique, reflection, questionnaires, observational studies	experiments concerning mental models		
Performer/ Composer	reflective practice, development of repertoire, long-term engagement (longitudinal study?)	quantitative methods for evaluation of user interface, mapping, etc.	quantitative methods for hardware/ software testing	
Designer	observation, questionnaire, Informal feedback	quantitative methods for user interface evaluation		use cases, feedback regarding stakeholder satisfaction
Manufacturer	market surveys, sales	sales, consumer feedback	quantitative methods for hardware/ software testing, consumer feedback	market penetration (performers, consumers), sales, consumer feedback

Note. Reprinted [adapted] from "A Framework and Tools for Mapping of Digital Musical Instruments," by Malloch, J., 2013, Retrieved from https://adobe.ly/2HNNKHS.

Additionally, Young (2016) further emphasized the third step of an evaluation is to consider "HCI paradigms that are relevant to computing for specific applications" (p. 99). As shown in table 4, Young (2016) proposed a framework derived from the field of HCI to "assess a musical device's functionality, usability, and the musician's overall user experience" (p. 99).

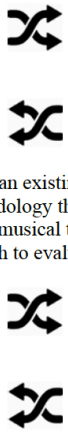
Table 4: A Framework of DMI Evaluation



Note. Reprinted [adapted] from "Human-Computer Interaction Methodologies Applied in the Evaluation of Haptic Digital Musical Intruments," by Young, G. W., 2016, Retrieved from https://adobe.ly/2HMBe2.

In terms of functionality, Wanderley and Orio (2002) proposed a simple task-based approach, underlining Fitts' law, Meyer's law, and Steering law specifically in the musical context (p. 3-5). More specifically, Wanderley and Orio described Fitt's law: "predicts that the time needed to point to a target of width *W* at a linear distance *A* away from the initial hand position is *T* seconds", Meyer's law enhanced Fitt's law "for the number of sub-movements can be created as many as possible", and Steering law can be applied "for evaluation of constrained motion for a generic curved path" (2002, p. 4-5). In addition to select the most suitable musical tasks, Wanderley and Orio highlighted a list of features used as guidelines for the evaluation of controllers' usability and the development of musical tasks (2002, p. 10). Wanderley and Orio described "Learningbility" is the time that required to learn how to use a device; "Explorability" allows users to discover the number of function and capabilities and understand how they are implemented; "Feature Controllability" is based on the constant changes of sound parameter; "Timing controllability" is "the fundamental difference between classical HCI observations and musical interactions is the central role of time" (2002, p. 10). After considering these feature, Wanderley and Orio (2002) suggested a list of musical tasks in order to evaluate the functionality of DMIs, such as selecting isolated tones and creating phrases with different contours (p. 11). More precisely, Young (2016) presented a list which includes a number of musical tasks linked with the evaluation techniques in HCI, as shown in table 5.

Table 5: Musical tasks and evaluation techniques from HCI

Musical Tasks		Existing HCI Functionality Evaluation Methodologies
Selecting an isolated tone: simple triggering to varying parameters such as pitch, loudness, and timbre. Musical gestures: glissandi, trills, grace notes, etc. Selecting scales and arpeggios at different speed, range, and articulation. Creating phrase contours: from monotonic to random. Ability to modulate timbre, amplitude or pitch for a given note and inside a phrase. Playing rhythms at different speeds and combining tones or pre-recorded materials. Synchronisation of musical processes.	 Select an existing HCI methodology that best fits the musical task you wish to evaluate	Target Acquisition - Fitts' Law. Pursuit Tracking - Control:Display ratio. Constrained Linear Motion Tracking. Constrained Circular Motion Tracking. Free-Hand Inking – subjective evaluation of facsimile signature. Aimed movements composed as sub-movements - Meyer's Law. Measuring trajectory movements - Steering Law. Circular motion path tracking and varying trajectories path tracking

Note. Reprinted [adapted] from "Human-Computer Interaction Methodologies Applied in the Evaluation of Haptic Digital Musical Intruments," by Young, G. W., 2016, Retrieved from https://adobe.ly/2HMBe2.

In terms of usability, Young (2016) emphasized that usability assessment "is used in HCI analyses to raise issues of efficiency, effectiveness, and user satisfaction" (p. 103). As shown in table 6, Young described a usability assessment technique called "NASA Task Load Index (NASA-TLX)" measures six subscales: "mental, physical, temporal, performance, effort, and frustration level" (2016, p. 104). In order to extrapolate information related to the factors of Learnability, Explorability, Feature Controllability, and Timing Controllability, Young (2016) discussed "the overall workload can be analyzed by using these subscales", and the data can be gathered by "using a modified post-task self-report or a Subject Mental Effort Questionnaire (SMEQ) and a Single Ease Question (SEQ)" (p. 104-105).

In terms of user experience, Young (2016) discussed "assessing a user's experience is a

relatively new and innovative area of investigation within the field of HCI" (p. 105). In order to measure user experience, Young suggested to use "a modified Consumer Product Questionnaire (CPQ)" and "a post-task User Experience Questionnaire (UEQ)" (2016, p. 105). Young described when participants perform musical tasks, their experience can be collected by conducting a "Critical Incidents Technique (CIT) analysis" (2016, p. 105).

Table 6: NASA-TLX Rating Scale

Subscale	Description
Mental	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical	How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal	How much time pressure did you feel due to the rate or pace at which the task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals?
Effort	How hard did you have to work to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Note. Reprinted [adapted] from "Human-Computer Interaction Methodologies Applied in the Evaluation of Haptic Digital Musical Intruments," by Young, G. W., 2016, Retrieved from https://adobe.ly/2HMBe2.

Quantitative Methods

In the past few year, recurrent calls were made to apply quantitative approaches. With references to substantiate this claim, Stowell et al. (2009) described a quantitative approach called "Turing Test": "in which the computer is required to imitate a human being in an interrogation" (p. 5). Stowell et al. aimed to evaluate B-Keeper system by implementing a three-

way comparison, "incorporating human, machine, and a third control condition using a steady accompaniment which remains at a fixed tempo dictated by the drummer" (2009, p. 9). For each test, Stowell et al. described the drummers firstly performed three times by giving "four steady beats of the kick drum to set the tempo and start, then plays along to an accompaniment track" (2009, p. 9). Next, Stowell et al. described a human tapper "taps the tempo on the keyboard, keeping time with the drummer, but only for one of the three times will this be altering the tempo of the accompaniment" (2009, p. 9). Particularly, Stowell et al. described each drummer took the test twice "consisting of the three randomly-selected trials, and playing to two different accompaniments" (2009, p. 9). Following this, Stowell et al. highlighted that "the quantitative results gained by asking for opinion measures and performance ratings" (2009, p. 9).

As results, Stowell et al. found that participants and observers had confusion "between the B-Keeper and the Human Tapper, but not between the B-Keeper and the Steady Tempo" (2009, p. 12). Most importantly, Stowell et al. emphasized that "B-Keeper achieves its aim of synchronizing a piece of music with the tempo variations of a live drummer, in a manner similar to that obtained if a human performs the synchronization" (2009, p. 12). However, Stowell et al. revealed that this approach "cannot be applied to the vocal timbre-mapping system, since for that there is no reference against which to compare" (2009, p. 12). Overall, Stowell et al. highlighted that Turing Test "derives useful numerical results, but at the cost of imposing a predetermined conceptual framework onto the interactive situation, including for example the concept of human likeness which can be attained to a greater or lesser degree" (2009, p. 12). Another interesting finding underlined by Stowell et al. was that this approach can be adapted for "audience-centered evaluation because the independent observer in our musical Turing Test case study takes the role of audience" (2009, p. 13).

Qualitative Methods

In the past few year, recurrent calls were made to apply qualitative approaches. With references to substantiate this claim, Ghamri et al. (2016) described "two distinct schools exist in qualitative research: content analysis from Social Sciences and discourse analysis from Linguistics" (p. 3).

In terms of content analysis, EI-Shimy and Coopertock (2016) described it as "operates on the principle of grounded theory, or the notion that hypotheses are contained within and can be induced from data collected during an experiment" (p. 8). More specifically, Ghamri et al. (2016) applied content analysis to evaluate a DMIs called the "Ballagumi". Ghamri et al. described the Ballagumi "consists of a flexible physical interface cast in silicone with embedded optical fibers that act as bend or pressure sensors" (2016, p. 3). Next, Ghamri et al. highlighted that the experiment was divided in two sessions including "the replication task, improvisation, and interviews" (2016, p. 5). Later, Ghamri et al. collected the participants' verbal descriptions during the interviews and applied content analysis to classify the verbal data into four main categories: "Instrument", "Learning", "Experimental Procedure" and "Mapping Differences" (2016, p. 5).

As results, Ghamri et al. (2016) described three main outcomes from the study. Firstly, the experimental method supported participants to concentrate on creating music instead of completing the tasks. Secondly, there are certain aspects of the Ballagumi such as "its weight and signal latency" should be improved (p. 8). For example, Ghamri et al. emphasized that "a DMI that is physically interactive, we may not need to explicitly add physicality to the mapping layer" because it makes the instrument less intuitive to play (2016, p. 8). Lastly, Ghamri et al. stated the experimental method presented in the paper "could be adapted and applied to evaluate any

technological device meant for musical expression" (2016, p. 9). More specifically, Ghamri et al. underlined that "by involving professional musicians, tasks and improvisation, it contributes to bridge the gap between research and artistic performance" (2016, p. 9).

In terms of discourse analysis, Stowell et al. (2009) described it as "an analytic tradition that provides a structured way to analyze the construction and reification of social structures in discourse" (p. 3). In the study, Stowell et al. intended to evaluate "the timbre remapping system with beatboxers", and five participants were recruited and participated in the solo and group sessions (2009, p. 4). More specifically, Stowell et al. (2009) described the solo sessions involved three steps, such as "free exploration, guided exploration, and semi-structured interview" (p. 4). Following this, peer group discussions were conducted in the experiment. Later, Stowell et al. (2009) applied the discourse analysis which consists of five steps. The speech data was transcribed, and then the analyst read it and took notes on "surface impressions and free associations". After itemising every single object in the discourse, the analyst reconstructed the described word and exam the context (p. 4).

As results, Stowell et al. (2009) emphasized that discourse analysis "can extract a detailed reconstruction of users' conceptualization of a system" (p. 12). Comparing with other methods such as observation or questionnaire, discourse analysis obtains their findings more clearly because Stowell et al. identified the discourse analysis provides "an interesting detail on the interaction between such concepts as controllability and randomness in the use of the interface, and the different ways of construing the interface itself" (2009, p. 12). However, Stowell et al. explained the group sessions were less structured, which "produced wide-ranging discourse with less content bearing directly on the interface" (2009, p. 12). In order to address this problem, Stowell et al. (2009) suggested the experimental contexts should be created to

encourage “exploration of the interface itself, while taking care not to lead participants in unduly influencing the categories and concepts they might use to conceptualize a system” (p. 12).

Mixed-Methods

In the past few year, recurrent calls were made to combine qualitative and quantitative approaches. With references to substantiate this claim, Kiefer et al. (2008) presented a usability study on the Nintendo Wiimote controller. Kiefer et al. relied on the mixed-methods: logged quantitative data combined with the analysis of the qualitative interview data to analyze “the gesture as well as the performance feedback from participants” (2008, p. 2).

More specifically, Kiefer et al. applied the simplistic musical tasks derived from Wanderley’s guidelines, such as “triggering (with drumming-like motions), continuous control using the roll and tilt axes, and gestural control using shape recognition” to evaluate the Wiimote’s use in more complex musical situations (2008, p. 2). In the meantime, Kiefer et al. provided the Roland HPD-15 Hand-Sonic for participants to compare with the Wiimote by using “triggering task, precise control task, expressive control task, and gestural recognition task” (2008, p. 2). Moreover, participants were interviewed after each task and asked about their experience of using the Willmote and their preferred controller. During each task and interview, the participants were videoed in order to observe their gestures. Following this, the qualitative interview data and quantitative log file data were analyzed. For the interview data, participants’ answers for each question were summarized and then transcribed the main answers and stored in the database. The next step was to code the quotes based on the categories and recode the quotes until the categorization was stable. The final step was to summarize the categorized quotes and produce the final results (Kiefer et al., 2008, p. 2). In addition to the quantitative data, Kiefer et

al. processed the log files in SuperCollider in order to "extract specific data such as timing information from the triggering task" (2008, p. 2).

As results, Kiefer et al. discussed the analysis of the qualitative interview data and logged quantitative data. The analysis of the interview data provided some unexpected issues in certain tasks and some suggestions on how to use the controllers. For instance, it was difficult for participants to determine the triggering point. And the participants felt it was unnatural "how going past certain points of rotation" (2008, p. 2). In the meantime, Kiefer et al. highlighted "the Wiimote's intuitive nature when used for expressive control: some participants explained it as embodied but others said that it widened the scope of editing possibilities" (2008, p. 2). Following this, Kiefer et al. stated these results demonstrated "the benefits of conducting a usability study, the kind of data that is difficult to determine purely by intuition alone and that is best collected from the observations of a larger group of people" (2008, p. 3). In addition, Kiefer et al. described the quantitative results "provided objective backup to certain elements of the interview results, some useful data about the functional side of the controller, and insight into global trends of the participants" (2008, p. 3). However, Kiefer et al. stated the results "lacked real-time data concerning the participants' experience of using the device due to lack of technology methodology" (2008, p. 4). In order to address this problem, Kiefer et al. described the third-paradigm HCI as "a growing trend in HCI research towards experience focused rather than task focused HCI" (2008, p. 4). More specifically, Kiefer et al. emphasized that the third-paradigm HCI has a potential impact for computer music because "the two fields share the common goal of evaluating experience and affect between technology and its users" (2008, p. 4).

8. Table of Dependent Variables

Description of DMI Evaluations

The first column contains all the dependent variables from the evaluation methods in each study. The second column describes what the evaluation methods were applied in each study. The third column presents what stakeholders were involved in each study. The fourth and fifth columns describe the conceptual and operation definition for each variable. The six column highlights how the data were collected. The seventh column provides a number linked with each reference. The last column shows what instruments were evaluated in each study.

Dependent Variable	Methods	Stakeholder	Conceptual Definition	Operational Definition	Data Collection	References	DMIs
Experience	Mixed	Performer	Their experience of using the controller	Interviews after each task	Script of interview questions and video recorded.	10	The Nintendo Wiimote and the Roland HandSonic
	Qualitative	Performer	Their experience of using the interface	A short semi-structured interview to discuss their experiences	Video recorded and interview transcript	11	Two interfaces for remapping timbre
	Qualitative	Performer	Their experience of playing the instrument	Interviews after each session. Questions: e.g. How would you describe your progression throughout the 3 tasks? How was your experience playing the instrument?	Audio and Videotaped	16	The Ballagumi (an alternative interface)
	Qualitative	Performer	Their experience of	Online Study: 7-Point-Likert	Questionnaire	24	Any type of musical

			testing an instrument	scale ranging from "Strongly disagree" to "Strongly agree"			instrument was accepted.
Enjoyment	Qualitative	Audience		Critique, reflection, questionnaires, observational studies		14	Any
	Qualitative	Performer		Reflective practice, development of repertoire, long-term engagement (longitudinal study?)		14	Any
	Qualitative	Designer		Observation, questionnaire, Informal feedback		14	Any
	Qualitative	Manufacture		market surveys, sales		14	Any
Playability	Quantitative	Audience		Experiments concerning mental models		14	Any
	Quantitative	Performer		Quantitative methods for evaluation of user interface, mapping, etc.		14	Any
	Quantitative	Designer		Quantitative methods for user interface evaluation		14	Any
	Quantitative	Manufacture		Sales, consumer feedback		14	Any

Robustness	Quantitative	Performer		Quantitative methods for hardware/ software testing		14	Any
	Quantitative	Manufacture		Quantitative methods for hardware/ software testing, consumer feedback		14	Any
Achievement of Design Specifications	Qualitative	Designer		Use cases, feedback regarding stakeholder satisfaction		14	Any
	Qualitative	Manufacture		market penetration (performers, consumers), sales, consumer feedback		14	Any
Mental	Qualitative	Performer	How mentally demanding was the task?	Numerical rating (1-20) after each task	Questionnaire	26	Any
Physical	Qualitative	Performer	How physically demanding was the task?	Numerical rating (1-20) after each task	Questionnaire	26	Any
Temporal	Qualitative	Performer	How hurried or rushed was the pace of the task?	Numerical rating (1-20) after each task	Questionnaire	26	Any
Performance	Qualitative	Performer	How successful were you in accomplishing what you were asked to do?	Numerical rating (1-20) after each task	Questionnaire	26	Any
Effort	Qualitative	Performer	How hard did you have to work to accomplish	Numerical rating (1-20) after each task	Questionnaire	26	Any

			your level of performance?				
Frustration	Qualitative	Performer	How insecure, discouraged, irritated, stressed, or annoyed were you?	Numerical rating (1-20) after each task	Questionnaire	26	Any
Playability	Mixed	Performer	To test core musical capabilities	Time needed for performers to be able to play a scale	Recording the different sessions and measuring the time	10	The Nintendo Wiimote and the Roland HandSonic
Playability	Quantitative	Performer	To assess real-time interaction	3-way switch (B-Keeper, Steady tempo, and Human tapping)	Questionnaire (Likert scale rating)	11	Beat-tracker

Description of HCI Evaluations

The first column contains all the dependent variables from the evaluation method in the study. The second column describes what evaluation method was used in the study. The third column presents what users were involved in the study. The fourth and fifth columns describe the conceptual and operation definition for each variable. The six column highlights how the data was collected. The seventh column provides a number linked with the reference. The last column shows what computer system was evaluated in the study.

Dependent Variable	Methods	Stakeholder	Conceptual Definition	Operational Definition	Data Collection	Reference	Computer System
Perceived Usability (PUs)	Quantitative (Task-based)	Students	Users' affective (e.g., frustration) and cognitive (e.g., effort)	7-point Likert scale from strongly disagree (1)	A modified version of WiIRE (Web Interactive Information	28	wikiSearch system

			responses to the system	to strongly agree (7); *questions shown in Appendix	Retrieval Experimentation) contained study instructions, the consent form, and demographic, pre-task, post-task and post-session questionnaires		
Aesthetics (AE)			The users' perception of the visual appearance of a computer application interface				
Novelty (NO)			Users' level of interest in the task and curiosity evoked by the system and its contents				
Felt Involvement (FI)			Users' feelings of being drawn in, interested, and having fun during the interaction				
Focused Attention (FA)			The concentration of mental activity; contained some elements of Flow, specifically focused concentration, absorption, and temporal dissociation				
Endurability (EN)			Users' overall evaluation of the experience, its perceived success and whether users would recommend the e-shopping site to others.				

			This factor combines concepts related to users' likelihood to return and evaluation of system success.				

9. Conclusion

Discussion and Recommendations

Young (2016) emphasized "a number of steps to ensure that a complete and in-depth device appraisal is carried out" and a framework to measure a "device's functionality, usability, and user experience" (p. 11). Most importantly, Young (2016) highlighted two main take away lessons which are "functionality testing should also include an element of analysis of the usability and user experience" and "usability and user experience in a musical context requires a longitudinal study as musicians must be given time to evaluate a device in a natural setting over time" (p. 11).

Additionally, Stowell et al. (2009) stated that "live music-making using interactive systems is not completely amenable to traditional HCI evaluation metrics" (p. 1). More specifically, Stowell et al. (2009) emphasized that "Talk-aloud" method is not practical in the musical performance, since it interrupts the process of music-making (p. 14). In order to solve this issue, Stowell et al. (2009) discussed Discourse Analysis and Turing Test which are both appropriate to evaluate DMIs. With a reference to substantiate Stowell et al.'s claim, EI-Shimy and Coopertock (2016) emphasized that Discourse Analysis represents "a strong social

constructionist attitude in which key categories and concepts are not predetermined but are considered an important outcome of the analysis" and Turing Test "derives useful numerical results, albeit at the expense of imposing a predetermined conceptual framework on the interaction" (p. 10).

In conclusion, the evaluation methods of computer interface in HCI are a well-established. However, none can be said to be fully adaptable with respects to DMIs. For example, the task-based approach proposed by Wanderley and Orio has limitations. More specifically, Kiefer et al. (2008) found that they could not capture "real-time" data about the user experience because "HCI methodology has evolved around a task-based paradigm and the stimulus response interaction model of WIMP systems, as opposed to the richer and more complex interactions that occur between musicians and machines" (p. 1). In order to address this challenge, Kiefer et al (2008) described third-paradigm HCI as "response to the evolving ways in which technology is utilized as computing becomes more increasingly embedded in daily life (p. 4). Most importantly, Kiefer et al. (2008) emphasized that third-wave HCI is "particularly suited to the design and evaluation of novel interactive musical interfaces" (p. 4). With a reference to substantiate Kiefer et al.'s claim, EI-Shimy and Coopertock (2016) highlighted that "the experience-based approach has become increasingly common among designers of interactive arts, musical interfaces, and playful systems keen on adopting a human-centered perspective" (p. 10).

Limitations of This Research

We did not cover all the papers from the reading list because we only considered the literatures relating to the evaluation methods of DMIs derived from the field of HCI. However,

additional papers could be reviewed, such as “Longitudinal Evaluation of the Integration of Digital Musical” from Gelineck and Serafin. This paper describes a longitudinal approach to the qualitative evaluation of DMIs focused on creativity and exploration. “Towards A New Conceptual Framework for Digital Musical Instruments” from Malloch, Birnbaum, Sinyor, and Wanderley. This paper describes the adaptation of an existing model of human information processing for the categorization of DMIs in terms of performance context and behavior.

Limitations of Previous Research

The vast majority of publications discussed the perspectives of performer and designer. However, it’s also important to include the audience’s perspective. The vast majority of papers rely on either qualitative and quantitative individually, or both. However, there remains a lack of mixed-methods. In addition, the vast majority of papers discussed how to measure user engagement in HCI. However, there remains a lack of evidence on the adaptation of existing assessment of user-driven techniques in HCI for the evaluation of DMIs in terms of user experience.

Future Direction of This Research

DMI designers are encouraged to tailor the evaluation techniques discussed in the paper to their own needs. More specifically, we reviewed the existing literatures on the suitable evaluation techniques. In order to inspire and guide designers, they are encouraged to consider the user-driven techniques discussed in the paper.

10. Appendix

Principal axis factoring with promax rotation of four-factor model.

	UES Sub-scale	Factor 1	Factor 2	Factor 3	Factor 4
I felt discouraged while using wikiSearch.	PUs	0.801	0.012	-0.023	0.017
I felt frustrated while using wikiSearch.	PUs	0.793	0.079	-0.095	-0.014
I felt annoyed with using wikiSearch.	PUs	0.739	0.119	-0.06	-0.026
This search experience did not work out the way I had planned.	EN	0.703	-0.047	0	0
I could not do some of the things I needed to do using wikiSearch.	PUs	0.682	-0.03	-0.065	0.006
I found wikiSearch confusing to use.	PUs	0.671	-0.132	0.117	0.03
Using wikiSearch was mentally taxing.	PUs	0.628	-0.165	0.059	0.043
This search experience was demanding.	PUs	0.594	-0.176	-0.02	0.039
I felt in control of the searching experience.	PUs	0.497	0.155	0.089	-0.058
I felt interested in my searching tasks.	NO	-0.073	0.81	-0.037	-0.037
The content of wikiSearch incited my curiosity.	NO	-0.153	0.81	-0.112	-0.01
My search experience was fun.	FI	-0.038	0.653	0.048	0.009
I felt involved in the searching tasks.	FI	-0.089	0.609	0.08	0.087
My search experience was rewarding.	EN	0.017	0.603	0.077	0.041
I would recommend wikiSearch to my friends and family.	EN	0.224	0.559	0.068	-0.114
I was really drawn into my searching tasks.	FI	0.011	0.526	0.029	0.284
I consider my search experience a success.	EN	0.372	0.486	-0.057	-0.014
Searching using wikiSearch was worthwhile.	EN	0.222	0.441	0.179	-0.008
The screen layout of wikiSearch appealed to my visual senses.	AE	0.046	-0.165	0.907	0.096
The screen layout of wikiSearch appealed to my visual senses.	AE	-0.013	-0.036	0.868	-0.015
The wikiSearch interface is aesthetically appealing.	AE	0.001	0.014	0.804	0.007
The wikiSearch interface is attractive.	AE	-0.019	0.122	0.791	-0.101
I liked the graphics and images used by wikiSearch.	AE	-0.074	0.162	0.509	-0.023
I was so involved in my searching task that I lost track of time.	FA	0.021	-0.055	0.011	0.77
The time I spent searching just slipped away.	FA	0.093	-0.055	-0.015	0.721
I lost myself in this searching experience.	FA	-0.077	0.087	-0.081	0.635
I blocked out things around me when I was using wikiSearch.	FA	0.016	-0.026	0.04	0.61
I was absorbed in my searching task.	FA	-0.026	0.28	0.035	0.503

11. Reference

1. Orio, N., Schnell, N., & Wanderley, M. M. (2001). Input Devices for Musical Expression: Borrowing Tools from HCI. Proceedings of the 2001 International Conference on New Interfaces for Musical Expression, 15–18. (https://adobe.ly/2Hwmxm1)

2. Wanderley, M., & Orio, N. (2002). Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI. Computer Music Journal, 26(3), 62– 76. (https://adobe.ly/2EZnbTP)

5. Jordà, S. (2004). Instruments and Players: Some Thoughts on Digital Lutherie. Journal of New Music Research, 33, 321–341. (https://adobe.ly/2EXRn1B)

7. Malloch, J., Birnbaum, D., Sinyor, E., & Wanderley, M. (2006). Towards a New Conceptual Framework for Digital Musical Instruments. In Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06) (pp. 49– 52). Montreal, Quebec, Canada. (https://adobe.ly/2vv0Aii)

9. Malloch, J. (2007, August). A Consort of Gestural Musical Controllers: Design, Construction, and Performance. McGill University, Montreal, Canada. (https://adobe.ly/2HySH05)

10. Kiefer, C., Collins, N., & Fitzpatrick, G. (2008). HCI Methodology For Evaluating Musical Controllers: A Case Study. In A. Camurri, G. Volpe, & S. Serafin (Eds.), Proceedings of the International Confer-ence on New Interfaces for Musical Expression (pp. 87–90). Genoa, Italy. (https://adobe.ly/2J6OptU)

11. Stowell, D., Robertson, A., Bryan-Kinns, N., & Plumbley, M. D. (2009). Evaluation of live human-computer music-making: Quantitative and qualitative approaches. International Journal of Human-Computer Studies, 67(11), 960–975. (https://adobe.ly/2F0SaP1)

14. O'Modhrain, S. (2011). A Framework for the Evaluation of Digital Musical Instruments. Computer Music Journal, 35(1), 28-42. (https://adobe.ly/2HdASjH)

17. Malloch, J. (2013). A Framework and Tools for Mapping of Digital Musical Instruments, (December). (https://adobe.ly/2HNNkHS)

19. Barbosa, J., Malloch, J., Huot, S., & Wanderley, M. (2015). What does "Evaluation" mean for the NIME community? New Interfaces for Musical Expression (NIME). (https://adobe.ly/2HLSBA1)

22. Young, G. W., & Murphy, D. (2015). HCI Models for Digital Musical Instruments: Methodologies for Rigorous Testing of Digital Musical Instruments. In International Symposium on Computer Mu-sic Multidisciplinary Research (CMMR). Plymouth, UK. (https://adobe.ly/2qIAHGL)

23. El-Shimy, D., & Cooperstock, J. R. (2016). User-Driven Techniques for the Design and Evaluation of New Musical Interfaces. Computer Music Journal, 40(2),35–46.https://doi.org/10.1162/COMJ_a_00357 (https://adobe.ly/2HMaKgO)

24. Schmid, G., Tuch, A. N., Papetti, S., & Opwis, K. (2016). Evaluating the Experiential Quality of Musical Instruments: A Psychometric Approach. In New Instruments for Musical Expression. (https://adobe.ly/2EYWZIU)

26. Young, G. W. (2016). Human-Computer Interaction Methodologies Applied in the Evaluation of Haptic Digital Musical Instruments. (https://adobe.ly/2HMBe2)

27. O'Brien, H., & Toms, E. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*,59(6), 938-955. doi:10.1002/asi.20801

28. O'Brien, H., & Toms, E. (2013). Examining the generalizability of the user engagement scale (ues) in exploratory search. *Information Processing and Management, 49*(5), 1092-1107. doi: 10.1016/j.ipm.2012.08.005

29. Dix, A., Finlay, J., Abowd, G. & Beale, R (2004). *Human-computer interaction* (3rd ed.). London: Pearson.